# Internship Proposal on theoretical online/reinforcement learning

–

## Internship description

Reinforcement Learning (RL) has become a fundamental paradigm for sequential decision-making under uncertainty, enabling agents to learn optimal behaviors through trial-and-error interactions with their environment. However, there remains a gap between theoretical analyses, which mostly focus on finite state and action spaces, and practical applications, which rely on large parametrized models.

The goal of this internship is to explore episodic reinforcement learning with parametric models. We will focus on the multinomial logistic function approximation of the transition kernel (see [11] for details). In this setting, the probability of transitioning from a state–action pair $(x_{n-1}, a_{n-1}) \in \mathcal{X} \times \mathcal{A}$ to a new state $x_n \in \mathcal{X}$ at time step $n \in \{1, \dots, N\}$ is modeled by a logistic model:

$$\mathbb{P}_n(x_n \mid x_{n-1}, a_{n-1}) = \frac{\exp\left(\phi(x_n \mid x_{n-1}, a_{n-1})^\top \theta_n\right)}{\sum_{x'_n} \exp\left(\phi(x'_n \mid x_{n-1}, a_{n-1})^\top \theta_n\right)}, \tag{1}$$

where $\phi$ is a known feature map and $\theta_n \in \mathbb{R}^d$ is an unknown parameter to be estimated as more data are collected. The feature map may be obtained, for instance, by omitting the final layer of a deep neural network that has been trained beforehand.

**Existing work and provisional research directions** Under this transition model, Hwang and Oh [8] and Li et al. [11] have designed efficient algorithms for episodic reinforcement learning, achieving regret upper bounds of order $O(d\sqrt{T})$, where $T$ is the number of episodes and $d$ is the dimension of the feature maps. In this internship, we aim to generalize their results by exploring one of the following possible research directions:

- *Nonparametric models:* Existing work has focused on finite-dimensional feature representations ($d < \infty$). When $d$ is infinite, both the computational complexity and the statistical guarantees (regret bounds) of existing algorithms become vacuous. We aim to study the case where the features belong to an infinite-dimensional reproducing kernel Hilbert space (RKHS). The goal is to leverage standard kernel techniques to replace $d$ with an effective dimension that captures the smoothness of the underlying function space (see Zenati et al. [18]).
- *Convex objective:* Let $\mu_n(x, a|\pi) \in \Delta_{\mathcal{X}\times\mathcal{A}}$ denote the probability of being in state–action pair $(x, a)$ when following policy $\pi$. Standard reinforcement learning maximizes the expected reward $\max_\pi \sum_{n=1}^N \langle \mu_n(\cdot|\pi), r_n \rangle$, where $r_n \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ are the reward vectors. In many applications, however, the objective is more general, as in the Concave Utility Reinforcement Learning (CURL) framework [7, 17], which seeks to minimize a convex function of the induced state–action distribution:

  $$\min_\pi f(\mu(\cdot|\pi)).$$

  Several machine learning problems can be cast as instances of CURL, including pure exploration [7, 13, 14], imitation and apprenticeship learning [5, 10, 16, 1], mean-field control [2], mean-field games with potential rewards [9], and risk-averse RL [3, 15, 6]. While RL has seen major progress in recent years, the theoretical understanding of CURL remains limited, mostly to finite state and action spaces. A promising direction for this internship is to study CURL under parametrized models—of either the policy or the transition dynamics—starting with the multinomial logistic approximation of the transition kernel.

An interesting application of CURL is demand-side management, which aims to control a large population of flexible electrical devices, such as water heaters, with the goal of making their aggregate electricity consumption follow a desired target. This application is particularly relevant for supporting the transition to renewable energy sources, whose production cannot be directly controlled. It is detailed in Moreno et al. [12] and could serve as a potential use case for applying the methods developed during the internship.

- *Robustness to misspecification:* We aim to consider cases where the true transition model does not exactly follow the logistic form (1) but can be well approximated by (1) up to some error $\epsilon$. Indeed, for many practical scenarios (for instance when the reinforcement learning algorithm comes from a deep neural architecture), the model will not follow exactly (1). This will involve defining a reasonable error model, studying how this error propagates through the Bellman operators and how this error can be estimated.

**Provisional Internship Plan** The internship will follow the provisional plan outlined below:

1. Literature review and problem modeling
2. Proposal and implementation of new algorithms. Application of the methods to simple synthetic examples (e.g., grid worlds in [4]) or to real-world problems (e.g., demand-side management in [12]).
3. Theoretical analysis of the proposed algorithms
4. Writing the internship report

The objective is to continue this work as part of a PhD.

## Internship Information

The internship (and potential PhD) will be supervised by Nicolas Gast (GHOST team, LIG/Inria Grenoble) and Pierre Gaillard (THOTH team, LJK/Inria Grenoble). This project is supported by the MIAI Cluster (Multidisciplinary Institute in Artificial intelligence).

**Location**: Inria (655 Av. de l'Europe, 38330 Montbonnot-Saint-Martin) and/or IMAG (150 Pl. du Torrent, 38400 Saint-Martin-d'Hères)
**Duration**: 4-6 months
**Starting Date**: March-May 2026
**Required knowledge**: Master's level or third-year engineering school.
**Profile**: machine learning, probability, statistics, optimization. Having completed a course in reinforcement learning or sequential learning (multi-armed bandits) is desirable.

## Contact


| | |
|---|---|
| Pierre Gaillard | Email: pierre.gaillard@inria.fr |
| Nicolas Gast | Email: nicolas.gast@inria.fr |


## References


[1] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning*, page 1, 2004.

[2] A. Bensoussan, P. Yam, and J. Frehse. *Mean Field Games and Mean Field Type Control Theory*. Springer, 2013.

[3] J. Garcıa and F. Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

[4] M. Geist, J. Pérolat, M. Laurière, R. Elie, S. Perrin, O. Bachem, R. Munos, and O. Pietquin. Concave utility reinforcement learning: The mean-field game viewpoint. *arXiv preprint arXiv:2106.03787*, 2021.

[5] S. K. S. Ghasemipour, R. Zemel, and S. Gu. A divergence minimization perspective on imitation learning methods. In *Conference on robot learning*, pages 1259–1277, 2020.

[6] I. Greenberg, Y. Chow, M. Ghavamzadeh, and S. Mannor. Efficient risk-averse reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 32639–32652, 2022.

[7] E. Hazan, S. Kakade, K. Singh, and A. Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pages 2681–2691, 2019.

[8] T. Hwang and M.-h. Oh. Model-based reinforcement learning with multinomial logistic function approximation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 7971–7979, 2023.

[9] P. Lavigne and L. Pfeiffer. Generalized conditional gradient and learning in potential mean field games. *Applied Mathematics & Optimization*, 88(3):89, 2023.

[10] J. W. Lavington, S. Vaswani, and M. Schmidt. Improved policy optimization for online imitation learning. In *Proceedings of The 1st Conference on Lifelong Learning Agents*, pages 1146–1173, 2022.

[11] L.-F. Li, Y.-J. Zhang, P. Zhao, and Z.-H. Zhou. Provably efficient reinforcement learning with multinomial logit function approximation. *Advances in Neural Information Processing Systems*, 37:58539–58573, 2024.

[12] B. M. Moreno, M. Brégère, P. Gaillard, and N. Oudjane. (Online) convex optimization for demand-side management: Application to thermostatically controlled loads. *Journal of Optimization Theory and Applications*, 205(3): 43, 2025.

[13] M. Mutti, L. Pratissoli, and M. Restelli. Task-agnostic exploration via policy gradient of a non-parametric state entropy estimate. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):9028–9036, 2021.

[14] M. Mutti, R. De Santi, and M. Restelli. The importance of non-Markovianity in maximum state entropy exploration. In *Proceedings of the 39th International Conference on Machine Learning*, pages 16223–16239, 2022.

[15] X. Pan, D. Seita, Y. Gao, and J. Canny. Risk averse robust adversarial reinforcement learning. In *International Conference on Robotics and Automation (ICRA)*, pages 8522–8528, 2019.

[16] T. Zahavy, A. Cohen, H. Kaplan, and Y. Mansour. Apprenticeship learning via Frank-Wolfe. In *AAAI Conference on Artificial Intelligence*, 2019. URL `https://api.semanticscholar.org/CorpusID:207869871`.

[17] T. Zahavy, B. O' Donoghue, G. Desjardins, and S. Singh. Reward is enough for convex MDPs. In *Advances in Neural Information Processing Systems*, pages 25746–25759, 2021.

[18] H. Zenati, A. Bietti, E. Diemert, J. Mairal, M. Martin, and P. Gaillard. Efficient kernelized ucb for contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 5689–5720. PMLR, 2022.